One Loss to Rule Them All: Marked Time-to-Event for Structured EHR Foundation Models

Anonymous Authors[1]

Abstract

Clinical events captured in Electronic Health Records (EHRs) are irregularly sampled and may consist of a mixture of discrete events and numerical measurements, such as laboratory values or treatment dosages. The sequential nature of EHR, analogous to natural language, has motivated the use of next-token prediction to train prior EHR Foundation Models (FMs) over events. However, this pre-training fails to capture the full structure of EHR. We propose *ORA* [1], a marked time-to-event pretraining objective that jointly models event timing and associated measurements. Across multiple datasets, downstream tasks, and model architectures, this objective consistently yields more generalizable representations than existing pretraining losses. Importantly, the proposed objective yields improvements beyond traditional classification evaluation, including better regression and time-to-event prediction. Beyond introducing a new FM, our results suggest a broader takeaway: pretraining objectives that account for all EHR dimensions are critical for expanding downstream capabilities and generalizability.

1. Introduction

Inspired by the success of large language models (LLMs), recent foundation models in Electronic Health Records (EHR) framed medical trajectories as sequences of discrete events typically pretrained using next-token prediction analogous to those used in natural language processing (Li et al., 2020; Pang et al., 2024; 2021; Odgaard et al., 2024). Despite these advances, EHR data fundamentally differ from natural language. EHR events occur at irregular intervals, multiple clinically relevant events may follow the same history, and many events are associated with numerical values such as laboratory measurements or dosages. Most prior work addresses these differences primarily through tokenization or introducing positional encoding (Pang et al., 2024; Wornow et al., 2024a; Hur et al., 2023). However, these modifications focus on how data should be represented as input and often overlook how to capture the underlying data-generating process in the EHR pretraining losses. The most common loss, next-token prediction, is a coarse surrogate for the underlying clinical process. Recent time-to-event formulations are a promising alternative (Steinberg et al., 2024; Gadd et al., 2025; Burger et al., 2025), but they often change the tokenizer, architecture, and loss simultaneously, making it difficult to isolate the effect of the pretraining loss in comparison to gains due to improved data representation (see full literature review in App. A). In this paper, we aim to answer: *Does a loss function accounting for EHR irregularity and numerical values yield more generalizable representations than the existing approaches?*

To answer this question, we introduce ORA, a marked time-to-event objective that, for each clinical code, models when it will next occur, and its associated value. We evaluate ORA with both Transformer and Mamba base architectures using the same tokenizer and comparable model size ($\sim$120M). Across two large EHR datasets, the proposed pretraining yields an average 11% gain across 15 downstream tasks spanning binary classification, time-to-event prediction, and regression.

Our contributions can be summarized as follows:

- **Novel loss.** Building on the parallel between EHR and marked point processes, we introduce ORA, a composite code-specific likelihood to capture EHR complexity.
- **Generalizability.** We demonstrate improved generalizability of ORA on top of both Mamba and Transformer base models across tasks and datasets.
- **Comprehensive downstream task evaluation.** Our experiments include 7 binary classification, 4 time-to-event prediction, and 4 regression tasks across two datasets. Our results demonstrate that ORA improves performance in diverse clinically-meaningful tasks.

2. ORA: EHR as marked point process

Notations. We consider each patient $i \in [1, 2, \cdots, N]$'s EHR data to consist of a tuple $\mathcal{H}_i := \{(t, m, v)_{i,j}, j \in$

[1]Anonymized code available at https://anonymous.4open.science/r/EHR_TTE/

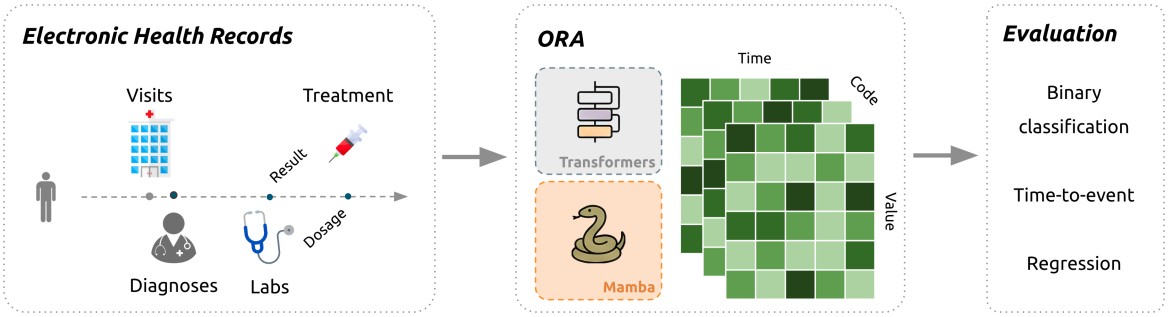

Figure 1. Our work introduces ORA, a marked time-to-event pretraining loss that accounts for the value and irregular timing of EHR, demonstrating the importance of the loss design for improved downstream capabilities and generalizability.

$[1, 2, \cdots, N_i]\}$ where $t \in \mathbb{R}^+$, is the timestamp of the event, $m \in \mathcal{M}$ the clinical code associated with the event, such as ICD-10 diagnosis or procedure codes, and $v \in \mathbb{R}$ the optional numerical value associated with the event. N_i denotes the number of events observed for patient i. For a patient i, events up to j (ordered by time) are denoted by $\mathcal{H}_{i,<j}$. That is, $\mathcal{H}_{i,<j} := \{(t, m, v)_{i,l}, l \in [1, 2, \cdots, j-1]\}$. Finally, we denote by $f^m_{i,j}$, the first occurrence of code m after the time associated with event j, and the set of the first events of all codes as $\mathcal{F}_{i,j} = \{\forall m, f^m_{i,j}\}$. Specifically, we define:

$$f^m_{i,j} = (\Delta t^m_j, v, \delta, m)$$

where, if an event of type m occurs after t_j, Δt^m_j corresponds to the interval between the current time t_j and the time to the next event m, and $\delta = 1$ as an event indicator. If no such event occurs in the future, we denote Δt^m_j as the interval up to the last observation, and set the event indicator to $\delta = 0$. In this context, δ denotes whether an event m is observed within the recorded EHR for patient i after the observation j or censored. Accounting for such unobserved events is critical, as ignoring them leads to biased likelihood estimates (Chen et al., 2024).

Joint likelihood. An intuitive approach is to maximize the joint likelihood of the full marked point process, corresponding to the joint between the time to the next event, the associated code, and value: $\mathcal{L}(\theta) = \prod_{i=1}^{N} p_\theta(\mathcal{H}_i) = \prod_{i=1}^{N} \prod_{j \in [1..N_i]} p_\theta((t, m, v)_{i,j} \mid \mathcal{H}_{i,<j})$

However, maximizing this joint likelihood poses some challenges. First, many downstream predictions depend on sets of future events and values rather than a single next token. For example, disease phenotyping is often characterized by multiple (potentially temporally ordered) codes, combined with conditions on their values (e.g., abnormal lab results). Simply optimizing the next-event loss may miss this multiplicity and long-term representativeness. Second, NTP provides sparse supervision because only the immediate next event contributes to the loss. Rare events often have limited impact on the overall likelihood. Third, this method implicitly treats future events as mutually exclusive even though multiple clinically meaningful events may co-occur.

ORA. To address these challenges, we instead propose to optimize the composite code-specific likelihood. At each observation j, we jointly model the first occurrence time and the value of every code:

$$\tilde{\mathcal{L}}_{\text{ORA}}(\theta) = \prod_i \prod_j \prod_{(\Delta t, m, v, \delta) \in \mathcal{F}_{i,j}} p_{\theta_m}(\Delta t, v, \delta \mid \mathcal{H}_{i,<j})$$

This replaces single-event supervision with dense multi-code supervision and naturally incorporates censoring. ORA assumes conditional independence across codes given the history; we make this trade-off to avoid imposing an arbitrary ordering over co-occurring future events.

Discretized implementation. The central challenge in computing the proposed likelihood is the estimation of the probability $p_{\theta_m}(\Delta t, v, \delta \mid \mathcal{H}_{i,<j})$. Previous work in temporal point analysis often constrain the intensity function to have a closed-form integral through parametric assumptions (Du et al., 2016; Mei & Eisner, 2017), which are computationally expensive. In this work, we follow Lee et al. (2018) to discretize time and value using code-specific quantiles. Suppose T and V denote the number of discretized bins for time and value. For each code m in the vocabulary, the model outputs a probability matrix $P^m_\theta[k, l] \in [0, 1]^{T \times V}$, representing the probability of observing an event in the k^{th} time-quantile and l^{th} value-quantile. Under this discretization, the log-likelihood can be expressed as:

$$\begin{aligned}\log \tilde{\mathcal{L}}_{\text{ORA}}(\theta) := \sum_{i=1}^{N} \sum_{j \in [1 \ldots N_i]} \sum_{(t,m,v,\delta) \in \mathcal{F}_{i,j}} \\ \delta \log P^m_\theta[q_m(t, v)](\mathcal{H}_{i,<j}) \\ +(1-\delta) \log \left(1 - \sum_{k=1}^{q_m(t,\cdot)} \sum_{l} P^m_\theta[k, l](\mathcal{H}_{i,<j})\right)\end{aligned}$$

For observed events, this objective is equal to the cross entropy loss by maximizing the probability mass $P^m_\theta[q_m(t, v)]$

in the associated time and value quantile where (t, v) belongs. For censored events, it enforces a low probability of observing any event before the end of the observation window, marginalizing over all possible values. The second term is essential in modeling the MTPP as ignoring censoring likelihood biases risk estimates.

3. Experiments

Datasets and tasks. We pretrain on MIMIC-IV and a private EHR dataset from a large urban hospital called 'Institution' hereafter. MIMIC-IV contains roughly 364K patients with mostly inpatient visits and ICU data. 'Institution Dataset' contains 6.7 million patients with both inpatient and outpatient visits. We evaluate 15 clinically meaningful linear-probe tasks: 7 binary classification tasks, 4 time-to-event tasks, and 4 regression tasks. Dataset splits and precise cohort definitions are in App. C.1.

- Binary Classification Tasks: In-hospital mortality (Mortality), Length-of-stay (LOS), Readmission, AMI, MASLD, Celiac and Stroke.
- Time-to-event tasks: AMI, MASLD, Celiac and Stroke.
- Regression tasks: Platelets, Creatinine, Oxygen (PaO2 or SpO2), and Glucose. These are laboratory measurements related to sepsis patients.

Baselines To demonstrate the effectiveness of our marked time to event loss function, we compare against two state-of-the-art EHR FMs: Context-Clues (Wornow et al., 2024a) and MOTOR (Steinberg et al., 2024). Context-Clues uses next-token prediction as its pretraining loss and includes architectures such as Llama (CC-Llama) and Mamba (CC-Mamba). MOTOR is a time-to-event FM that predicts the time to the next event. These two FMs have been shown to achieve the best performance, as reported in (Pang et al., 2025). We also compare against task-specific baselines: XGBoost, DeepHit, and most-recent-value imputation. These are strong baselines used in recent studies on EHR foundation models (Steinberg et al., 2021; 2024; Im et al., 2025).

ORA Implementation We follow Steinberg et al. (2024) by using an entropy filter to construct the vocabulary and joint encoding to synthesize different clinical concepts (e.g., medical codes, numerical values, etc.). To verify that our loss function is robust across different architectures, we used both Transformer (Vaswani et al., 2017) and Mamba (Gu & Dao, 2024) as backbones. For a fair comparison, we set the parameter size for all models to be around 120M. Full model descriptions are detailed in App. B.

Evaluation Once pretrained, we freeze the FMs and fit task-specific linear heads: logistic regression for classification, DeepHit for time-to-event prediction, and linear regression for numerical forecasting on the extracted embeddings. We report AUROC, time-dependent C-index, and R^2.

4. Results

We pretrain all models on MIMIC-IV and Institution. Then we evaluate them on 7 classification tasks, 4 time-to-event tasks, and 4 regression tasks. Note that Celiac is only evaluated in Institution data. We present the results for MIMIC-IV in the following sections. ORA denotes the best-performing model between ORA-Transformer and ORA-Mamba. The full results can be found in App. D.

Table 1. AUROC of Binary Classification Tasks for MIMIC-IV Patients. *Larger is better.*

Model	Readmission	LOS	Mortality	AMI	MASLD	Stroke
XGBoost	0.726 (0.003)	0.748 (0.003)	0.882 (0.005)	0.807 (0.008)	0.700 (0.020)	0.708 (0.023)
CC-Llama	0.730 (0.003)	0.781 (0.003)	0.901 (0.005)	0.796 (0.010)	0.663 (0.020)	0.696 (0.025)
CC-Mamba	0.731 (0.003)	0.779 (0.003)	0.896 (0.006)	0.783 (0.009)	0.661 (0.019)	0.665 (0.024)
MOTOR	0.743 (0.002)	0.833 (0.002)	0.954 (0.0023)	0.825 (0.008)	0.703 (0.021)	0.710 (0.019)
ORA	**0.747 (0.003)**	**0.841 (0.002)**	**0.965 (0.002)**	**0.832 (0.007)**	**0.722 (0.020)**	**0.711 (0.018)**
Rel. Improv.	+0.54%	+0.96%	+1.15%	+0.85%	+2.70%	+0.14%

Table 2. C-index of Time-to-event Phenotype Tasks for MIMIC-IV Patients. *Larger is better.*

Model	AMI	MASLD	Stroke
Deephit	0.651 (0.029)	0.560 (0.024)	0.574 (0.085)
CC-Llama	0.782 (0.030)	0.687 (0.023)	0.847 (0.039)
CC-Mamba	0.760 (0.035)	0.664 (0.021)	0.845 (0.039)
MOTOR	0.798 (0.027)	0.696 (0.018)	0.828 (0.035)
ORA	**0.847 (0.029)**	**0.735 (0.016)**	**0.899 (0.024)**
Rel. Improv.	+6.14%	+5.60%	+6.14%

ORA improves classification. Across MIMIC-IV and Institution datasets, ORA improves over Context-Clues baselines on every classification task with most gains between 2% and 5%. However, the improvement over MOTOR (pre-trained using time-to-event loss) is small (0% to 2%), indicating that compared with time-to-event loss, predicting time and value jointly is not necessarily needed for downstream binary classification tasks.

ORA is strongest on time-to-event and regression. ORA improves time-to-event prediction over all baselines, and yields an average relative gain of $5\% - 7\%$. The clearest benefit appears on regression: ORA consistently beats

Table 3. R^2 of Lab Test Regression for MIMIC-IV Patients. *Larger is better.* * means its improvement is only calculated with respect to the best FM baseline (except the most-recent-value imputation).

Model	Creatinine	PaO2	Platelets	Glucose
Most Recent	**0.877 (0.034)**	-0.431 (0.024)	**0.910 (0.002)**	-0.030 (0.035)
CC-Llama	0.319 (0.018)	0.219 (0.005)	0.290 (0.004)	0.155 (0.015)
CC-Mamba	0.307 (0.017)	0.218 (0.005)	0.261 (0.004)	0.148 (0.014)
MOTOR	0.556 (0.033)	0.231 (0.005)	0.298 (0.004)	0.163 (0.016)
ORA	0.645 (0.034)	**0.286 (0.006)**	0.659 (0.003)	**0.186 (0.018)**
Rel. Improv.	16.00% *	+23.81%	+121.14%*	+14.11%

both Context-Clues and MOTOR baselines across all regression tasks, with task-wise improvements ranging from 16% to 121%. This suggests that explicitly modeling event-associated values during pretraining transfers most directly to numerical forecasting. The most recent value imputation has divergent performance in different tasks, which we discuss in Section 5.

ORA is robust across different architectures. To exclusively study the influence of the marked time-to-event loss function in ORA, we design additional ablation studies by fixing the tokenizer of all models. For both Transformer and Mamba architecture, we only vary the pretraining loss function: Next-token Prediction (NTP), Temporal Point Process (TPP), and ORA. NTP only predicts the next code. TPP predicts both the time and code. ORA jointly predicts the time and value for all codes. As shown in Appendix D, ORA outperforms NTP and is on par with TPP in most binary classification and time-to-event tasks, and shows the largest gain (up to 121%) in regression tasks. Across both backbones, the central conclusion is consistent: pretraining objectives aligned with irregular timing and values produce more generalizable EHR representations.

5. Discussion

Our work introduces ORA, a novel pretraining loss that accounts for temporal irregularity and numerical values associated with EHR events. While previous works have proposed approaches to account for these dimensions through tokenization and the choice of pretraining losses (Burger et al., 2025; Gadd et al., 2025), no work isolates the sources of relative improvements across diverse and clinically-meaningful downstream tasks. Our core contribution is to demonstrate that the choice of pretraining loss alone is responsible for a critical improvement in the generalizability of FMs. The current literature on FMs often focuses on novel architectures, with design choices driven by performance. Isolating the impact of these choices has seldom been studied, yet it is critical for further advancement in EHR FMs. Our contribution aims to advance the mathematical foundation of these models and, consequently, inform the development of future ones. Furthermore, we extend linear-probe evaluation beyond classification to include time-to-event modeling and regression. ORA, which captures EHR temporal irregularities and their associated values in the pretraining loss, improves performance across the broadest range of downstream tasks, regardless of model architecture, across multiple datasets.

Robust performance in lab test regression Table 3 suggests that the most recent value imputation baseline is particularly strong for lab measurements with high short-term stability, such as creatinine and platelets. If a lab changes slowly, the latest observed value is already a strong predictor. In contrast, ORA performs best on measurements that quickly evolve, such as PaO2 and glucose. To quantify this, we measured the median relative change between consecutive observations: creatinine and platelets change by only about 9%, whereas PaO2 and glucose change by roughly 14–18%. These larger short-term fluctuations make last-value imputation less effective, consistent with the strongly negative R^2 for PaO2 and glucose. This is also clinically plausible, as PaO2 and glucose can change rapidly in response to treatment and acute physiologic status. Finally, ORA consistently outperforms the other foundation model baselines across all four regression tasks, indicating better overall robustness across heterogeneous lab outcomes.

Interpreting ORA in clinical contexts. By training models on both the timing and measurement of laboratory measurements, our loss function have the potential to closely approximate the clinical practice, where $60-70\%$ of diagnoses are made using laboratory tests (Agarwal, 2014). Unlike other parts of the structured EHR (e.g., diagnoses, prescriptions, procedures) that represent healthcare processes, laboratory measurements provide unique biological insights into a person at the time they are tested. Thus, improving our capabilities to simulate and forecast laboratory measurements in conjunction with vital signs and diagnoses has important implications for holistic modeling of patient trajectories (Renc et al., 2024).

Limitations and future work. Our evaluation focuses on generalizability of different tasks. Given prior evidence that multi-task training improves robustness under distribution shift (Jeanselme et al., 2025), an important direction for future work is to assess ORA's pretraining using multiple datasets. Meanwhile, we fix the model capacity at approximately 120M parameters. While we demonstrate that ORA outperforms NTP in most tasks, future work should examine whether the observed improvements persist as we increase the model parameters.

Impact Statement

Foundation models in healthcare promise to democratize access to advanced modeling capabilities for institutions lacking the resources to build models from scratch. Yet this promise comes with risks to privacy, fairness, and safety. Although this paper advances the mathematical foundations of these models, such risks must be rigorously addressed before any clinical implementation.

References

Agarwal, R. Quality-improvement measures as effective ways of preventing laboratory errors. *Lab. Med.*, 45(2): e80–e88, May 2014.

Alaa, A. M., Hu, S., and Schaar, M. Learning from clinical judgments: Semi-markov-modulated marked hawkes processes for risk prognosis. In *International conference on machine learning*, pp. 60–69. PMLR, 2017.

An, U., Lee, S. A., Jeong, M., Gorla, A., Chiang, J. N., and Sankararaman, S. Dk-behrt: Teaching language models international classification of disease (icd) codes using known disease descriptions. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pp. 133–143. PMLR, 2025.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Bhave, S. and Perotte, A. Point processes for competing observations with recurrent networks (popcorn): A generative model of ehr data. In *Machine Learning for Healthcare Conference*, pp. 770–789. PMLR, 2021.

Burger, M., Chopard, D., Londschien, M., Sergeev, F., Yèche, H., Kuznetsova, R., Faltys, M., Gerdes, E., Leshetkina, P., Bühlmann, P., et al. A foundation model for intensive care: Unlocking generalization across tasks and domains at scale. *medRxiv*, pp. 2025–07, 2025.

Chang, C., Wang, W.-Y., Peng, W.-C., and Chen, T.-F. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20, 2025a.

Chang, Y., Boyd, A. J., Xiao, C., Kass-Hout, T., Bhatia, P., Smyth, P., and Warrington, A. Deep continuous-time state-space models for marked event sequences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Chen, G. H. et al. An introduction to deep survival analysis models for predicting time-to-event outcomes. *Foundations and Trends® in Machine Learning*, 17(6):921–1100, 2024.

Cui, H., Unell, A., Chen, B., Fries, J. A., Alsentzer, E., Koyejo, S., and Shah, N. H. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *NPJ Digital Medicine*, 8(1):577, 2025.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1555–1564, 2016.

Enguehard, J., Busbridge, D., Bozson, A., Woodcock, C., and Hammerla, N. Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*, pp. 85–113. PMLR, 2020.

Fallahpour, A., Alinoori, M., Ye, W., Cao, X., Afkanpour, A., and Krishnan, A. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

Gadd, C., Gokhale, K., Acharya, A., Cooper, J., Crowe, F., Fitzsimmons, L., Jackson, T., Nirantharakumar, K., Yau, C., and collaborative, O. Survivehr: a competing risks, time-to-event foundation model for multiple long-term conditions from primary care electronic health records. *medRxiv*, pp. 2025–08, 2025.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Hegselmann, S., von Arnim, G., Rheude, T., Kronenberg, N., Sontag, D., Hindricks, G., Eils, R., and Wild, B. Large language models are powerful electronic health record encoders. *arXiv preprint arXiv:2502.17403*, 2025.

Hill, B. L., Emami, M., Nori, V. S., Cordova-Palomera, A., Tillman, R. E., and Halperin, E. Chiron: a generative foundation model for structured sequential medical data. 2023.

Hripcsak, G., Shang, N., Peissig, P. L., Rasmussen, L. V., Liu, C., Benoit, B., Carroll, R. J., Carrell, D. S., Denny, J. C., Dikilitas, O., et al. Facilitating phenotype transfer using a common data model. *Journal of biomedical informatics*, 96:103253, 2019.

Hur, K., Oh, J., Kim, J., Kim, J., Lee, M. J., Cho, E., Moon, S.-E., Kim, Y.-H., Atallah, L., and Choi, E. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2023.

Im, S., Oh, J., and Choi, E. Labtop: A unified model for lab test outcome prediction on electronic health records. *arXiv preprint arXiv:2502.14259*, 2025.

Islam, K. T., Shelton, C. R., Casse, J. I., and Wetzel, R. Marked point process for severity of illness assessment. In *Machine learning for healthcare conference*, pp. 255–270. PMLR, 2017.

Jeanselme, V. *Clinical Presence: Impact on Predictive Modelling and Algorithmic Fairness*. PhD thesis, University of Cambridge (United Kingdom), 2024.

Jeanselme, V., Martin, G., Sperrin, M., Peek, N., Tom, B., and Barrett, J. Prediction of survival outcomes under clinical presence shift: A joint neural network architecture. *arXiv preprint arXiv:2508.05472*, 2025.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023.

Karami, H., Atienza, D., and Ionescu, A. Tee4ehr: Transformer event encoder for better representation learning in electronic health records. *Artificial Intelligence in Medicine*, 154:102903, 2024.

Lee, C., Zame, W., Yoon, J., and Van Der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Lee, S. A., Jain, S., Chen, A., Ono, K., Fang, J., Rudas, A., and Chiang, J. N. Emergency department decision support using clinical pseudo-notes. *arXiv preprint arXiv:2402.00160*, 2024.

Lee, S. A., Jain, S., Chen, A., Ono, K., Biswas, A., Rudas, Á., Fang, J., and Chiang, J. N. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1):394, 2025.

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

McDermott, M., Nestor, B., Argaw, P., and Kohane, I. S. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems*, 36:24322–24334, 2023.

Mei, H. and Eisner, J. M. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Odgaard, M., Klein, K. V., Thysen, S. M., Jimenez-Solem, E., Sillesen, M., and Nielsen, M. Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv preprint arXiv:2404.15201*, 2024.

Pang, C., Jiang, X., Kalluri, K. S., Spotnitz, M., Chen, R., Perotte, A., and Natarajan, K. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In Roy, S., Pfohl, S., Rocheteau, E., Tadesse, G. A., Oala, L., Falck, F., Zhou, Y., Shen, L., Zamzmi, G., Mugambi, P., Zirikly, A., McDermott, M. B. A., and Alsentzer, E. (eds.), *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 239–260. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/pang21a.html.

Pang, C., Jiang, X., Pavinkurve, N. P., Kalluri, K. S., Minto, E. L., Patterson, J., Zhang, L., Hripcsak, G., Gürsoy, G., Elhadad, N., et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

Pang, C., Jeanselme, V., Choi, Y. S., Jiang, X., Jing, Z., Kashyap, A., Kobayashi, Y., Li, Y., Pollet, F., Natarajan,

K., et al. Fomoh: A clinically meaningful foundation model evaluation for structured electronic health records. *arXiv preprint arXiv:2505.16941*, 2025.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Ren, W., Zhu, J., Liu, Z., Zhao, T., and Honavar, V. A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models. *arXiv preprint arXiv:2507.12774*, 2025.

Renc, P., Jia, Y., Samir, A. E., Was, J., Li, Q., Bates, D. W., and Sitek, A. Zero shot health trajectory prediction using transformer. *NPJ Digit. Med.*, 7(1):256, September 2024.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30, 2017.

Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., and Gerstung, M. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

Snyder, D. L. and Miller, M. I. *Random point processes in time and space*. Springer Science & Business Media, 2012.

Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., and Shah, N. H. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113: 103637, 2021.

Steinberg, E., Xu, Y., Fries, J. A., and Shah, N. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

Su, X., Messica, S., Huang, Y., Johnson, R., Fesser, L., Gao, S., Sahneh, F., and Zitnik, M. Multimodal medical code tokenizer. *arXiv preprint arXiv:2502.04397*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., and Shah, N. H. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023.

Wornow, M., Bedi, S., Hernandez, M. A. F., Steinberg, E., Fries, J. A., Ré, C., Koyejo, S., and Shah, N. H. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024a.

Wornow, M., Thapa, R., Steinberg, E., Fries, J., and Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36, 2024b.

A. Related Work

This section reviews FMs for structured EHR data, with an emphasis on how the literature adapts natural language strategies to model clinical data. We then examine the training objectives inherited from this field and discuss alternatives.

Foundation models for structured EHR. Echoing the success of language models, FMs have emerged as an effective paradigm in structured EHR modeling (Wornow et al., 2023; Steinberg et al., 2021; Odgaard et al., 2024; Pang et al., 2025; Cui et al., 2025; Steinberg et al., 2024; Wornow et al., 2024a; Gadd et al., 2025; Burger et al., 2025; Shmatko et al., 2025; Wornow et al., 2024b). Instead of task-specific models, FMs are pretrained on large amounts of structured EHR to extract representations with the goal of linear-probe predictive capabilities (Pang et al., 2025). We refer the reader to Ren et al. (2025) for an extensive survey of structure and unstructured EHR FMs, and focus this review on how the key differences between EHR from natural language.

EHR data is characteristically marked by irregular time points and may include numerical values, such as the time of measurement and the corresponding result. Ignoring these dimensions discards important proxies of patients' health (Jeanselme, 2024), potentially reducing the representativeness of FMs.

EHR representation. Two key alternatives have been proposed to represent these different dimensions in a format amenable to training FMs: serialization and tokenization. First, the availability of text-based general-purpose LLMs has spurred work in training representations using text-serialized EHR data (Su et al., 2025; Hegselmann et al., 2025; 2023; Lee et al., 2024; 2025; Cui et al., 2025). While our findings may provide insights for training such models, we focus on explicit tokenization of structured EHR elements, which represents the sequence of diagnosis codes, procedures, prescriptions, labs, and visit information as tokens using clinically relevant vocabularies (Hripcsak et al., 2019), as opposed to natural language vocabularies. For instance, BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), CEHR-BERT (Pang et al., 2021) and variants (An et al., 2025; Shmatko et al., 2025) leverage such token sequences to train BERT architectures (Devlin et al., 2019). More recently, decoder-based approaches reframe EHR modeling using next-token prediction (Pang et al., 2024). Beyond standard transformer architectures, state-space models like Mamba (Gu & Dao, 2023) have also been used to model EHR sequences, and demonstrate strong generalizability across different downstream tasks (Fallahpour et al., 2024; Wornow et al., 2024a).

However, the challenge associated with time and value remains in this tokenization. To handle irregular medical events, CEHR-BERT (Pang et al., 2021) introduces artificial time tokens as input. MOTOR (Steinberg et al., 2024) uses rotary position embeddings to integrate time in the attention mechanism. For numerical value encoding, Wornow et al. (2024a) discretizes numerical code into tokens. Alternatively, Hill et al. (2023) assigns measurement to a unique embedding.

An inherited pretraining loss from natural language. The parallel between medical event streams and words in a sentence has not only influenced the way to represent EHR but also to train such models (McDermott et al., 2023). Particularly, FMs' training often relies on maximizing the likelihood of the next token. Even in the more general context of time series, FMs often minimize this loss (Ansari et al., 2024) or the mean square error of the next value (Jin et al., 2023; Goswami et al., 2024; Chang et al., 2025a; Das et al., 2024), implicitly assuming temporal regularity in observation sequences. Such losses do not account for the temporal irregularities and associated values characteristic of EHR, a problem that has largely been overlooked in the development of FMs.

Marked point process. In statistics, irregularly sampled time series are typically modeled using temporal point processes. When events are associated with values, a marked point process captures the underlying stochastic process (Snyder & Miller, 2012; Daley & Vere-Jones, 2003). Prior work has explored improving predictive models by modeling EHR event types jointly with their occurrence times, often through point-process formulations (Du et al., 2016; Enguehard et al., 2020; Bhave & Perotte, 2021; Islam et al., 2017; Schulam & Saria, 2017; Alaa et al., 2017). More recent FMs extend this idea by introducing time-to-event pretraining objectives that require predicting not only which event occurs next, but also when it occurs (Karami et al., 2024; Steinberg et al., 2024; Chang et al., 2025b; Shmatko et al., 2025). Most closely related to our work, a couple of FMs model event-associated values (Gadd et al., 2025; Burger et al., 2025). However, the variability in tokenization strategies and architectural differences has made it challenging to isolate the impact of the choice of pretraining loss on downstream performance. Prior benchmarking evaluates performance gains in linear-probe classification tasks, demonstrating inconsistent gains and limited transportability across clinically meaningful predictions and models (Pang et al., 2025; Wornow et al., 2024b). In contrast, our work isolates the role of the pretraining objective, while expanding downstream evaluation to linear-probe regression and time-to-event prediction.

B. Implementing ORA

To maximize the previous likelihood, one must extract a representation of the medical history $\mathcal{H}_{i,<j}$ and compute the matrix $P^m(x)$. Importantly, the proposed pre-training loss is architecture agnostic. We propose evaluating its efficacy on common FMs architectures: an attention-based architecture (Transformer (Vaswani et al., 2017)) and a state-space model (Mamba (Gu & Dao, 2024)) to demonstrate its utility for modeling EHR data.

B.1. EHR Tokenization

To isolate the effects of the loss on performance, we fix the tokenizer to the one developed by Steinberg et al. (2024) and adopt a similar entropy-based filter to construct the vocabulary with the most informative events. Instead of taking the most frequent events in the vocabulary, we select codes with the highest entropy over the whole dataset. Define $p(m)$ as the probability that the code appears in each patient. The entropy calculation is as follows:

$$H(m) = -p(m)\log(m)$$

If the dataset has an ontology mapping, we can calculate the conditional entropy of any code m relative to its parent n. Suppose $p(m, n^+)$ denotes the probability that both m and n appear per patient and $p(m, n^-)$ be the probability that only m appears in the patient. $p(m) = p(m, n^+) + p(m, n^-)$, the conditional entropy is as follows:

$$H(m|n) = -p(m, n^-)\log\frac{p(m, n^-)}{p(m)} - p(m, n^+)\log\frac{p(m, n^+)}{p(m)}$$

B.2. Model Backbones

Following Steinberg et al. (2024), we adopt a decoder-only Transformer backbone that avoids look-ahead from future medical events through a causal attention mechanism. Unlike the standard Transformer, it uses rotary position embeddings with age for improved temporality processing. It also adopts local attention and sample packing for efficient pretraining.

As an alternative, we train Mamba, a state-space model introduced in Gu & Dao (2024), using the proposed pretraining loss. Mamba replaces attention-based modules in the Transformer with a selective state-space block, whose compute time scales linearly with sequence length. Prior work has shown its advantage over the Transformer in capturing long, irregular EHR sequences with temporal dependencies (Wornow et al., 2024a; Fallahpour et al., 2024).

For a fair comparison across different architectures, we set the parameter size for all models to around 120M. Within each architecture, we also use the same configuration files for different losses, which is specified as follows:

Table 4. Model configurations.

Model	Configuration	Value
Transformer	context length	8192
	learning rate	1e−5
	dim model	768
	intermedicate size	3072
	num layers	11
	num heads	12
	Total Parameters	**119M**
Mamba	context length	8192
	learning rate	2e−4
	dim model	768
	intermedicate size	1536
	num layers	28
	num heads	16
	Total Parameters	**120M**

B.3. Efficient Projection Head

The last layer of our architectures includes a head to project the output embedding $E \in \mathbb{R}^D$, with D is the model's hidden dimension, to the probability matrix $\forall m \in \mathcal{M}, P^m(x) \in (0,1)^{T \times V}$. A direct projection from $\mathbb{R}^D$ to the discretized joint of shape $T \times V \times |\mathcal{M}|$ would require an impractical number of parameters to estimate. Instead, following (Steinberg et al., 2024), we use a factorized two-stage computation. First, a one-layer fully connected neural network projects the embedding E to time-specific features $H_j \in \mathbb{R}^{T \times D_2}$. Then we use a second projection followed by a final Softmax for all codes with numerical values, projecting H_j into the final matrix $P^m \in [0,1]^{T \times V}$. Similarly, for nonnumerical codes, a single-layer project into $P^m \in [0,1]^{T \times 1}$.

In our experiments, we use $D = 768$, $D_2 = 512$, $T = 8$, $V = 10$. $|\mathcal{M}|$ is decided by the vocabulary size of each dataset. This factorization reduces the number of parameters by 20% compared to a fully connected network that projects the embedding onto the discretized joint. The visualization is as follows. M_{non} and M_{num} represent nonnumerical codes and numerical codes for pretraining, respectively.

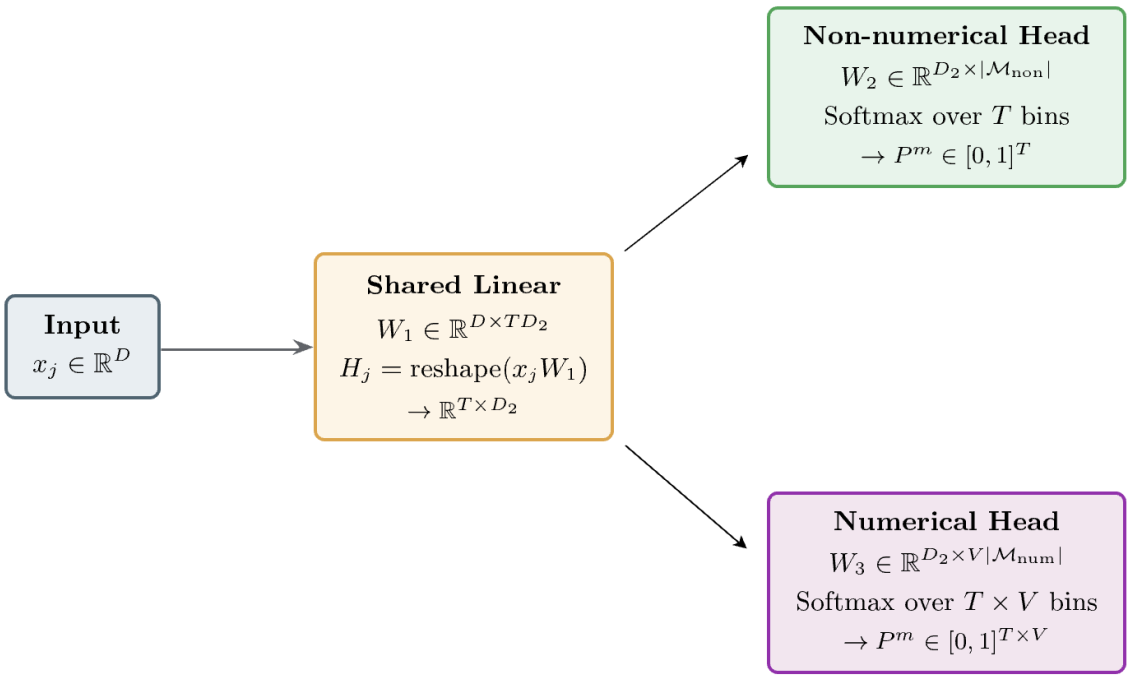

Figure 2. Efficient prediction head: We share the same projection layer for both nonnumerical and numerical codes and then use two different projection heads to output the separate probability matrices.

C. Experimental setting

C.1. Datasets

Table 5. Dataset Split of MIMIC-IV and Institution

Split Name	MIMIC-IV		Institution	
	# Patients	# Events	# Patients	# Events
Training Set	291,702	579,667,440	3,996,578	1,562,316,866
Tuning Set	36463	71,789,378	705,279	273,247,565
Test Set	36462	71,290,747	2,015,082	781,462,164

C.2. Cohort Definition

C.2.1. Outcome Definition

We construct three outcome tasks using the same definitions as in (Pang et al., 2025). The detailed inclusion criteria can be found in section 3.2 of the referenced paper.

C.2.2. Phenotype Definition

For each disease, we define a set of at-risk events as cohort inclusion criteria and a set of case events to determine patients' labels. Compared with using a set of ICD codes, this method adds more task difficulty and is more clinically meaningful.

C.2.3. Regression Definition

We define regression tasks as predicting the lab test after 4 hours of the prediction time. Each lab event can be identified with a corresponding code in the following table:

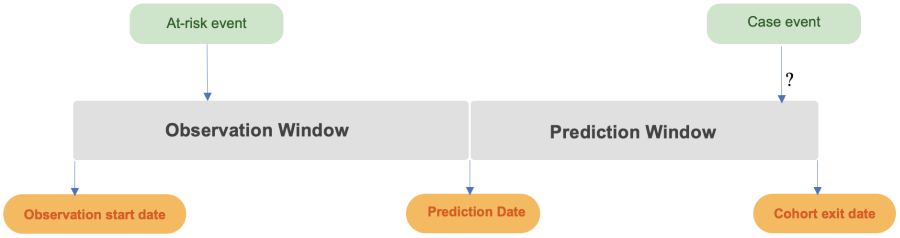

Figure 3. Visualization of Cohort Definition. The detailed definition of at-risk, prediction time, and case events can be found in Appendix B of (Pang et al., 2025).

Table 6. Corresponding Medical Codes for All Regression Tasks on MIMIC and Institution

Task	MIMIV-IV	Institution(OMOP)
Creatinine	50912	LOINC/2160-0
Platelets	51265	LOINC/26515-7
Oxygen	50821 (PaO_2)	LOINC/2708-6 (SpO_2)

D. Experiment Results

This section details the performance for each architecture, loss, and task across both datasets.

D.1. MIMIC-IV

We evaluate all models on 6 classification tasks, 3 time-to-event tasks and 3 regression tasks on MIMIC-IV.

D.1.1. Classification

Table 7. AUROC of Binary Classification Tasks for MIMIC-IV Patients. *Larger is better.*

Model	Readmission	LOS	Mortality	AMI	MASLD	Stroke
XGBoost	0.726 (0.003)	0.748 (0.003)	0.882 (0.005)	0.807 (0.008)	0.700 (0.020)	0.708 (0.023)
CC-Llama	0.730 (0.003)	0.781 (0.003)	0.901 (0.005)	0.796 (0.010)	0.663 (0.020)	0.696 (0.025)
CC-Mamba	0.731 (0.003)	0.779 (0.003)	0.896 (0.006)	0.783 (0.009)	0.661 (0.019)	0.665 (0.024)
MOTOR	0.743 (0.002)	0.833 (0.002)	0.954 (0.0023)	0.825 (0.008)	0.703 (0.021)	0.710 (0.019)
Transformer-NTP	0.728 (0.003)	0.808 (0.002)	0.926 (0.004)	0.810 (0.008)	0.691 (0.020)	0.691 (0.025)
Transformer-TPP	0.744 (0.004)	0.839 (0.002)	0.963 (0.003)	0.818 (0.007)	0.705 (0.018)	0.700 (0.020)
Transformer-ORA	0.745 (0.003)	**0.841 (0.002)**	**0.965 (0.002)**	0.827 (0.007)	0.714 (0.017)	0.711 (0.018)
Mamba-NTP	0.732 (0.004)	0.813 (0.002)	0.933 (0.004)	0.814 (0.008)	0.679 (0.020)	**0.719 (0.022)**
Mamba-TPP	0.741 (0.003)	0.809 (0.003)	0.934 (0.004)	0.822 (0.007)	0.709 (0.020)	0.666 (0.017)
Mamba-ORA	**0.747 (0.003)**	0.812 (0.002)	0.939 (0.004)	**0.832 (0.007)**	**0.722 (0.020)**	0.660 (0.022)

D.1.2. TIME-TO-EVENT

Table 8. C-index of Time-to-event Phenotype Tasks for MIMIC-IV Patients. *Larger is better.*

Model	AMI	MASLD	Stroke
Deephit	0.651 (0.029)	0.560 (0.024)	0.574 (0.085)
CC-Llama	0.782 (0.030)	0.687 (0.023)	0.847 (0.039)
CC-Mamba	0.760 (0.035)	0.664 (0.021)	0.845 (0.039)
MOTOR	0.798 (0.027)	0.696 (0.018)	0.828 (0.035)
Transformer-NTP	0.760 (0.034)	0.679 (0.018)	0.767 (0.064)
Transformer-TPP	0.839 (0.028)	0.702 (0.019)	**0.899 (0.023)**
Transformer-Numerical	0.823 (0.025)	0.658 (0.019)	0.821 (0.039)
Transformer-ORA	0.819 (0.029)	0.735 (0.016)	**0.899 (0.024)**
Mamba-NTP	0.773 (0.030)	0.667 (0.018)	0.743 (0.064)
Mamba-TPP	0.844 (0.029)	**0.739 (0.017)**	0.825 (0.034)
Mamba-ORA	**0.847 (0.029)**	0.703 (0.019)	0.837 (0.034)

D.1.3. REGRESSION

Table 9. R^2 of Lab Test Regression for MIMIC-IV Patients. *Larger coefficient reflects a larger proportion of variance explained.*

Model	Creatinine	PaO2	Platelets
CC-Llama	0.319 (0.018)	0.219 (0.005)	0.290 (0.004)
CC-Mamba	0.307 (0.017)	0.218 (0.005)	0.261 (0.004)
MOTOR	0.556 (0.033)	0.231 (0.005)	0.298 (0.004)
Most Recent	**0.877 (0.034)**	-0.431 (0.024)	**0.910 (0.002)**
Transformer-NTP	0.489 (0.025)	0.259 (0.005)	0.312 (0.004)
Transformer-TPP	0.603 (0.030)	0.234 (0.005)	0.310 (0.003)
Transformer-ORA	0.603 (0.028)	0.267 (0.005)	0.605 (0.003)
Mamba-NTP	0.595 (0.029)	0.270 (0.005)	0.407 (0.003)
Mamba-TPP	0.617 (0.030)	0.241 (0.005)	0.400 (0.004)
Mamba-ORA	0.645 (0.034)	**0.286 (0.006)**	0.659 (0.003)

Table 10. RMSE of Lab Test Regression for MIMIC-IV Patients. *Lower is better.*

Model	Creatinine	PaO2	Platelets
Baseline	**0.503 (0.084)**	77.727 (0.535)	**38.632 (0.326)**
CC-Llama	1.244 (0.053)	57.369 (0.466)	108.897 (0.653)
CC-Mamba	1.255 (0.052)	57.421 (0.462)	111.113 (0.635)
MOTOR	0.968 (0.067)	56.920 (0.447)	108.285 (0.630)
Transformer-NTP	1.038 (0.058)	55.888 (0.461)	107.198 (0.647)
Transformer-TPP	0.915 (0.063)	56.816 (0.452)	107.359 (0.605)
Transformer-ORA	0.915 (0.062)	55.576 (0.439)	81.202 (0.593)
Mamba-NTP	0.925 (0.062)	55.475 (0.451)	99.533 (0.632)
Mamba-TPP	0.899 (0.064)	56.549 (0.447)	100.106 (0.567)
Mamba-ORA	0.866 (0.068)	**54.877 (0.429)**	75.476 (0.581)

Table 11. MAE of Lab Test Regression for MIMIC-IV Patients. *Lower is better.*

Model	Creatinine	PaO2	Platelets
Baseline	**0.193**	45.537	**25.184**
	(0.002)	(0.324)	**(0.136)**
CC-Llama	0.803	38.620	78.139
	(0.004)	(0.230)	(0.312)
CC-Mamba	0.811	38.593	80.145
	(0.004)	(0.219)	(0.317)
MOTOR	0.542	38.211	77.215
	(0.004)	(0.206)	(0.322)
Transformer-NTP	0.651	37.102	76.247
	(0.004)	(0.218)	(0.310)
Transformer-TPP	0.507	38.114	76.565
	(0.003)	(0.217)	(0.310)
Transformer-ORA	0.504	36.859	56.163
	(0.003)	(0.209)	(0.249)
Mamba-NTP	0.547	36.983	70.657
	(0.003)	(0.209)	(0.284)
Mamba-TPP	0.492	37.667	70.825
	(0.003)	(0.207)	(0.277)
Mamba-ORA	0.448	**36.319**	51.451
	(0.003)	**(0.199)**	(0.236)

D.2. Institution

As external validation, this section shows similar results across tasks and architectures on the Institution dataset, a large urban center.

D.2.1. CLASSIFICATION

Table 12. AUROC of Binary Classification Tasks for Institution Patients. *Larger is better.*

Model	Readmission	LOS	Mortality	AMI	MASLD	Stroke	Celiac
XGBoost	0.732 (0.003)	0.773 (0.002)	0.875 (0.005)	0.833 (0.008)	0.681 (0.009)	0.872 (0.006)	0.650 (0.031)
CC-Llama	0.758 (0.003)	0.798 (0.002)	0.892 (0.004)	0.821 (0.009)	0.731 (0.008)	0.846 (0.007)	0.604 (0.040)
CC-Mamba	0.757 (0.003)	0.784 (0.002)	0.883 (0.004)	0.816 (0.009)	0.706 (0.008)	0.865 (0.007)	0.635 (0.038)
MOTOR	0.783 (0.003)	0.850 (0.002)	0.963 (0.002)	0.853 (0.007)	0.727 (0.007)	0.869 (0.006)	0.647 (0.030)
Transformer-NTP	0.760 (0.003)	0.842 (0.002)	0.938 (0.003)	0.821 (0.008)	0.713 (0.009)	0.849 (0.007)	0.624 (0.039)
Transformer-TPP	0.784 (0.003)	0.857 (0.002)	0.963 (0.002)	0.843 (0.008)	0.718 (0.008)	0.863 (0.006)	0.736 (0.033)
Transformer-ORA	**0.787 (0.003)**	**0.866 (0.002)**	**0.968 (0.002)**	0.850 (0.008)	0.745 (0.007)	0.870 (0.007)	**0.749 (0.036)**
Mamba-NTP	0.757 (0.003)	0.817 (0.002)	0.915 (0.003)	0.780 (0.010)	0.681 (0.008)	0.830 (0.007)	0.637 (0.033)
Mamba-TPP	0.776 (0.003)	0.859 (0.002)	0.942 (0.003)	0.854 (0.008)	0.737 (0.008)	0.864 (0.006)	0.739 (0.034)
Mamba-ORA	0.779 (0.002)	0.861 (0.002)	0.946 (0.003)	**0.858 (0.008)**	**0.766 (0.008)**	**0.873 (0.006)**	0.744 (0.038)

D.2.2. TIME-TO-EVENT

Table 13. C-index of Time-to-event Phenotype Tasks for Institution Patients. *Larger is better.*

Model	AMI	Celiac	Ischemic Stroke	MASLD
Deephit	0.617 (0.009)	0.564 (0.029)	0.626 (0.009)	0.601 (0.005)
CC-Llama	0.719 (0.007)	0.581 (0.022)	**0.780 (0.006)**	0.620 (0.006)
CC-Mamba	0.732 (0.008)	0.673 (0.018)	0.773 (0.006)	0.603 (0.006)
MOTOR	**0.752 (0.007)**	0.593 (0.023)	0.760 (0.007)	0.624 (0.005)
Transformer-NTP	0.683 (0.007)	0.582 (0.026)	0.727 (0.008)	0.605 (0.006)
Transformer-TPP	0.729 (0.008)	0.577 (0.023)	0.748 (0.007)	0.616 (0.005)
Transformer-ORA	0.721 (0.009)	0.604 (0.020)	0.769 (0.006)	0.626 (0.005)
Mamba-NTP	0.693 (0.008)	0.519 (0.024)	0.721 (0.008)	0.613 (0.005)
Mamba-TPP	0.739 (0.008)	0.641 (0.023)	0.758 (0.008)	0.603 (0.005)
Mamba-ORA	0.747 (0.007)	**0.679 (0.021)**	0.767 (0.007)	**0.637 (0.006)**

D.2.3. REGRESSION

Table 14. R^2 of Lab Test Regression for Institution Patients. *Larger coefficient reflects a larger proportion of variance explained.*

Model	Creatinine	Platelets	SpO2
Most Recent	**0.679**	0.426	0.078
	(0.036)	(0.011)	(0.013)
CC-Llama	0.298	0.150	0.770
	(0.013)	(0.005)	(0.007)
CC-Mamba	0.277	0.146	0.762
	(0.013)	(0.006)	(0.007)
MOTOR	0.469	0.198	0.842
	(0.020)	(0.007)	(0.006)
Transformer-NTP	0.431	0.228	0.816
	(0.022)	(0.006)	(0.007)
Transformer-TPP	0.397	0.203	0.832
	(0.018)	(0.007)	(0.006)
Transformer-ORA	0.542	0.349	0.836
	(0.024)	(0.008)	(0.006)
Mamba-NTP	0.423	0.220	0.809
	(0.024)	(0.007)	(0.006)
Mamba-TPP	0.543	0.311	0.849
	(0.027)	(0.008)	(0.006)
Mamba-ORA	0.582	**0.464**	**0.852**
	(0.029)	**(0.007)**	**(0.006)**

Table 15. RMSE of Lab Test Regression for Institution Patients. *Lower is better.*

Model	Creatinine	Platelets	SpO2
Most Recent	**0.630**	74.419	15.598
	(0.047)	(1.048)	(0.225)
CC-Llama	0.987	90.486	7.802
	(0.035)	(0.941)	(0.114)
CC-Mamba	1.001	90.695	7.939
	(0.035)	(0.939)	(0.112)
MOTOR	0.825	88.308	6.437
	(0.042)	(0.948)	(0.123)
Transformer-NTP	0.856	86.837	6.948
	(0.039)	(0.969)	(0.110)
Transformer-TPP	0.879	88.046	6.638
	(0.041)	(0.936)	(0.117)
Transformer-ORA	0.766	79.574	6.555
	(0.044)	(0.883)	(0.122)
Mamba-NTP	0.860	87.097	7.078
	(0.045)	(0.881)	(0.110)
Mamba-TPP	0.765	81.869	6.301
	(0.047)	(0.911)	(0.123)
Mamba-ORA	0.732	**72.188**	**6.226**
	(0.047)	**(0.878)**	**(0.118)**

Table 16. MAE of Lab Test Regression for Institution Patients. *Lower is better.*

Model	Creatinine	Platelets	SpO2
Most Recent	**0.235 (0.004)**	**47.591 (0.428)**	6.377 (0.112)
CC-Llama	0.567 (0.006)	65.073 (0.458)	4.670 (0.046)
CC-Mamba	0.567 (0.006)	65.327 (0.488)	4.742 (0.045)
MOTOR	0.402 (0.006)	63.290 (0.478)	3.325 (0.038)
Transformer-NTP	0.445 (0.005)	62.540 (0.436)	3.968 (0.037)
Transformer-TPP	0.406 (0.006)	63.068 (0.435)	3.448 (0.037)
Transformer-ORA	0.341 (0.005)	56.882 (0.376)	3.377 (0.040)
Mamba-NTP	0.434 (0.006)	62.555 (0.393)	4.086 (0.039)
Mamba-TPP	0.340 (0.005)	58.465 (0.407)	3.196 (0.038)
Mamba-ORA	0.326 (0.005)	49.793 (0.357)	**3.103 (0.037)**